AI技术在档案智能分类与标引中的应用研究

赵 越
河南港投航程城建集团有限公司 河南 郑州 451100

摘 要：在数字时代国家发展战略的推动下，档案数字化转型加速，海量异构档案资源的高效管理成档案工作核心痛点。传统人工分类与标引模式效率低、标引不一致、主观性强，难以满足智能化需求。本文基于相关理论，梳理档案分类与标引内涵，分析传统模式局限及AI技术优势，探讨核心AI技术应用机理，构建系统架构，阐述模块功能与实现路径，结合实践验证技术有效性，为档案管理智能化转型提供理论与实践指引，助力挖掘档案价值、提升管理利用效能。

关键词：人工智能；档案管理；智能分类；自动标引；深度学习

引言：档案是记录社会发展、传承历史的核心信息资产，分类与标引是档案管理基础环节，影响检索、利用和价值挖掘效率。数字化转型使档案形态向海量、多态、异构演变，传统人工模式受专业水平和主观认知影响，存在效率低、标引不规范、检索不精准等问题，难以管理海量档案。人工智能技术，特别是自然语言处理和深度学习在文本处理领域的突破，为档案智能分类与标引提供革命性方案。本文聚焦AI应用，探索实现路径，推动档案管理智能化发展。

1 档案智能分类与标引的理论基础

1.1 档案分类与标引的基本概念

档案分类与标引是档案管理的基础，二者紧密配合，助力档案资源有序管理与高效利用。档案分类依据形成规律、内容特征等属性，遵循原则与标准，把无序档案划分成不同层级类别，让档案系统条理化，为后续工作筑牢根基。档案标引在分类基础上，提取核心主题与关键信息，赋予规范标识，分主题与分类标引两类，实现内容与检索需求精准对接。按全国团体标准定义，档案智能分类是借助 AI 技术自动或半自动识别特征并归类的过程，自动标引则是提取核心信息并规范标识，二者是档案智能管理的关键环节。

1.2 传统档案分类标引的方法与局限

传统档案分类标引以人工为主，常用分类法和主题标引法。分类法借助标准化体系，由专业人员判断类目完成分类编码；主题标引法人工提取核心词，结合词表规范处理。但传统方法问题不少：人工处理速度慢，面对海量档案力不从心；专业人员知识储备和主观判断不同，导致标引不一致，影响检索精准度；需大量专业人力，培训周期长，复杂体系培训需3 - 6个月，管理成本高；刚性层级结构难适应新兴领域档案，人工标引也跟不上档案内容多样化、复杂化发展[1]。

1.3 AI技术应用于档案管理的可行性与优势

AI技术应用于档案智能分类与标引可行且优势明显。从可行性上，档案数据特征与AI处理能力契合，自然语言处理可理解语义，机器学习能优化模型，破解人工瓶颈。且档案分类标引规则重复，适合AI标准化处理，现有标准也为其应用提供依据。从优势看，AI能实现自动化处理，速度是人工数十倍，大幅缩短周期；通过模型训练和算法优化，减少主观误差，部分自动标引模型准确率超 85%，分类模型超 90%；减少人工投入，缓解专业人员短缺；模型持续学习，适配新兴档案类型，实现动态优化，满足利用需求。

2 AI 核心技术及其在档案分类标引中的应用机理

2.1 自然语言处理技术

自然语言处理（NLP）是AI技术在档案分类标引中的核心支撑，其核心功能是实现计算机对人类语言的理解、分析与处理，破解档案文本的语义识别难题，为分类标引提供基础。在档案处理场景中，自然语言处理技术主要应用于文本预处理、语义分析、关键词提取三个环节。文本预处理阶段，通过分词、停用词去除、词形还原等操作，清理档案文本中的噪声数据，如超文本标记语言标签、特殊符号等，统一文本格式，结合档案领域专用词典提高分词准确性；语义分析阶段，通过词向量模型（如Word2Vec、GloVe）将文本词汇映射到低维向量空间，实现语义相似度计算，同时利用LDA主题模型识别档案潜在主题，如“人事任免”“项目审批”等；关键词提取阶段，结合TF-IDF算法与TextRank算法，计算词汇在档案文本中的重要性，筛选核心关键词，为自动标引提供支撑，其核心机理是将非结构化的档案文本转化为计算机可识别、可处理的结构化信息，实现档案

内容的深度解析。

2.2 机器学习与深度学习方法

机器学习与深度学习是实现档案智能分类与标引的核心算法支撑，二者通过数据训练实现模型优化，提升分类标引的精准度。机器学习方法主要包括监督学习、无监督学习两类，在档案分类标引中应用广泛：监督学习（如支持向量机、决策树）通过已标注的档案样本训练模型，学习档案特征与分类标引规则，实现新档案的自动分类与标引；无监督学习（如K-means聚类算法）无需人工标注样本，通过挖掘档案文本的内在特征，实现档案的自动聚类与分类，适用于未标注档案的批量处理[2]。深度学习作为机器学习的延伸，通过构建多层神经网络，实现档案特征的深度提取，其核心优势是能够处理复杂、高维度的档案数据，如多模态档案、长文本档案。常用的深度学习模型包括CNN（卷积神经网络）、BERT预训练模型等，其中BERT模型可实现上下文语义的精准理解，有效提升长文本档案分类标引的精准度，解决传统算法难以处理的语义歧义问题。

2.3 智能分类的技术路径

AI技术在档案智能分类中的应用遵循“数据输入—预处理—特征提取—模型训练—分类输出”的核心技术路径，形成完整的闭环流程。首先，数据输入阶段，收集各类档案数据，包括文本、图像、音视频等多模态档案，完成数据的整理与汇总，确保数据的完整性；其次，数据预处理阶段，对档案数据进行格式标准化、噪声去除、文本分词等操作，图像类档案需通过OCR技术转换为文本，其中印刷体OCR识别准确率不低于99.5%，清晰度较高的手写体不低于85%；再次，特征提取阶段，通过自然语言处理技术提取档案的文本特征、元数据特征、上下文特征等，结合词向量模型、TF-IDF算法将特征转化为模型可识别的向量；然后，模型训练阶段，选用合适的机器学习或深度学习模型，利用已标注的档案样本进行训练，通过调整模型参数优化训练效果，确保分类准确率；最后，分类输出阶段，将预处理后的新档案输入训练好的模型，模型自动识别档案特征，依据预定义的分类体系输出分类结果，同时保留人工审核环节，对分类结果进行校验与修正，确保分类的准确性与规范性。

2.4 智能标引的技术路径

档案智能标引的技术路径与智能分类相互关联、协同推进，核心是通过AI技术实现标引词的自动提取、规范化匹配与优化，具体分为四个步骤。第一步，核心信息提取，利用自然语言处理技术，结合TF-IDF与TextRank算法，提取档案文本中的关键词、主题词，挖掘档案的核心内容与特征，筛选出权重较高的初步标引词；第二步，标引词规范化，引入《中国档案主题词表》等标准词表，将初步提取的标引词与标准词表进行匹配，通过余弦相似度计算实现语义映射，对非标准标引词进行规范化处理，确保标引的一致性；第三步，标引模型训练，利用机器学习模型，结合已标注的档案样本，学习标引词与档案内容的对应关系，优化标引模型的精准度，减少标引误差；第四步，标引输出与优化，模型自动为档案赋予规范化的标引词与分类号，形成完整的标引结果，同时结合人工审核，对错误标引进行修正，通过持续的模型迭代，提升标引的精准度与效率，实现标引工作的自动化、规范化[3]。

3 档案智能分类与标引的系统架构与实现

3.1 总体架构设计

档案智能分类与标引系统采用分层解耦的微服务架构，遵循“辅助人工、提升效能、确保准确、保障安全”的原则，整体分为四层，从下至上依次为数据资源层、智能引擎层、应用服务层、用户交互层，各层相互支撑、协同运行，确保系统的稳定性、可扩展性与实用性。数据资源层是系统的基础，负责存储各类档案数据（文本、图像、音视频等）、标准词表、训练样本、模型参数等，采用分布式文件系统与混合数据库架构，支持海量数据的高效存储与访问；智能引擎层是系统的核心，封装自然语言处理模块、机器学习/深度学习模型、特征提取模块等，以标准化API接口提供智能分类、自动标引等核心服务；应用服务层封装档案管理的业务逻辑，包括任务调度、权限管理、日志审计、结果审核等，实现智能分类标引与传统档案管理业务的融合；用户交互层为档案管理员提供Web端、移动端等多渠道交互界面，支持数据上传、模型操作、结果查看与修正等功能，系统整体可用性不低于99.5%，平均无故障时间大于10000小时。

3.2 数据预处理模块

数据预处理模块是系统实现智能分类与标引的前提，核心功能是将原始档案数据转换为适合模型处理的标准格式，去除噪声、提升数据质量，主要包括四个子模块。数据采集与导入子模块，支持多种格式档案数据的批量导入，包括PDF、DOCX、JPG、MP4等，实现与现有档案管理系统的平滑对接，确保数据导入的便捷性；格式标准化子模块，对导入的档案数据进行格式统一，文本档案统一编码为UTF-8，图像、音视频档案进行格式转换，确保数据的一致性；文本清洗子模块，通

过正则表达式去除文本中的无关符号、错误字符，纠正OCR识别误差，分段分句，同时去除“的”“了”等停用词，结合档案领域专用词典进行分词与词形还原，增强文本规范性；数据校验子模块，对预处理后的数据进行校验，筛选出不完整、不合格的数据，提醒管理员进行补充与修正，确保数据质量，为后续模型训练与分类标引提供可靠支撑。

3.3 模型训练与优化

模型训练与优化模块是提升系统分类标引精准度的核心，负责模型的训练、参数调整、性能评估与迭代优化，主要包括三个核心环节。一是训练样本构建，收集不同类型、不同领域的档案数据，由专业人员进行人工标注，形成高质量的训练数据集与测试数据集，训练集需覆盖所有目标类目，每个类目的样本量满足模型训练需求，确保样本的代表性与完整性；二是模型训练与参数调整，根据档案数据特征，选用合适的模型（如BERT、支持向量机），利用训练数据集进行模型训练，通过调整学习率、迭代次数、窗口大小等参数，优化模型性能，减少过拟合、欠拟合问题，同时采用交叉验证方法，提升模型的泛化能力；三是模型评估与迭代，利用测试数据集对训练好的模型进行性能评估，采用准确率、召回率、F1值等核心指标进行衡量，当模型性能未达到预设标准时，通过增加训练样本、优化算法参数等方式进行迭代优化，确保智能分类模型准确率不低于90%，自动标引模型准确率不低于85%，满足档案管理的实际需求[4]。

3.4 系统功能模块

系统功能模块基于总体架构，结合档案管理的实际需求，实现智能分类、自动标引、结果审核、检索查询等核心功能，具体分为五大子模块。（1）智能分类子模块，支持批量与单份档案的自动分类，可根据预设的分类体系（如《中国档案分类法》）自动划分档案类目，生成分类编码，同时支持分类规则的自定义设置，适配不同行业、不同单位的档案分类需求；（2）自动标引子模块，实现档案关键词、主题词的自动提取与规范化标引，生成标引词列表与标引报告，支持标引词的手动修改与补充；（3）结果审核子模块，提供人工审核界面，管理员可对系统自动生成的分类与标引结果进行校验、修改与确认，确保结果的准确性与规范性，尤其在复杂价值鉴定、敏感信息识别等领域强化人工审核；（4）检索查询子模块，基于智能分类与标引结果，支持关键词检索、主题检索、分类检索等多种检索方式，实现档案的快速、精准检索，简单检索请求平均响应时间不超过2秒；（5）系统管理子模块，负责用户权限管理、日志管理、数据备份与恢复、模型更新等功能，保障系统的安全稳定运行，支持至少100个并发用户的在线操作。

结束语

AI技术应用于档案智能分类与标引，是档案事业数字化转型的关键突破口。本文从理论入手，梳理档案分类标引概念与传统方法局限，阐述自然语言处理等 AI 核心技术的应用机理，设计四层系统架构，明确三大模块内容。AI 能大幅提升分类准确率、标引效率，保证标引一致性与可扩展性，构建人机协同模式。未来，随着相关技术发展，档案智能分类标引将向更深层次演进，档案管理部门应推动AI与业务融合，助力事业高质量发展。

参考文献

[1]张若松.AI技术赋能声像档案智能化创新应用研究[J].机电兵船档案,2025(3):31-32,45.

[2]陈远霞.大数据时代下的机关档案智能分类与检索技术研究[J].办公自动化,2024,29(14):60-62.

[3]颜雅玲.基于人工智能的档案分类与标引自动化研究[J].办公自动化,2026,31(3):83-85.

[4]杨玺.人工智能赋能档案收管用:智能分类、检索与利用策略研究[J].机电兵船档案,2025(5):166-168.